Sampling strategies

Collecting Evidence

http://www.ck12.org/user3AAnoka/book/Anoka-Hennepin-Probability-and-Statistics/r56/section/4.2/Data-Collection-253A253Aof253A253A-Anoka-Hennepin-Probability-and-Statistics/

Collect Evidence

- Determine data collection method
- Find relevant information about problem in literature

Data collection methods

- Census. A census is a study that obtains data from every member of a population.
- Sample survey. A sample survey is a study that obtains data from a subset of a population.
- Experiment. An experiment is a controlled study to understand cause-and-effect relationships.

  In the analysis phase, the researcher compares group scores on some dependent variable. Based on the analysis, the researcher draws a conclusion about whether the treatment (independent variable) had a causal effect on the dependent variable.
- Observational study. Like experiments, observational studies attempt to understand cause-and-effect relationships. However, unlike experiments, the researcher is not able to control (1) how subjects are assigned to groups and/or (2) which treatments each group receives.

Advantages and disadvantages of various data collection techniques

- Resources. When the population is large, a sample survey has a big resource advantage over a census.
- Generalizability. Generalizability refers to the appropriateness of applying findings from a study to a larger population. Generalizability requires random selection.

  Observational studies do not feature random selection; so it is not appropriate to generalize from the results of an observational study to a larger population.
- Causal inference. Cause-and-effect relationships can be teased out when subjects are randomly assigned to groups. Therefore, experiments, which allow the researcher to control assignment of subjects to treatment groups, are the best method for investigating causal relationships.

Population - experiment

- Consider the research problem/question and define the subjects from which evidence should be collected, i.e. population
- Determine the site or location of the experiment to be representative of the defined population
- Experimental material to be used in the experiment should also be representative of the defined population

Population - survey

- Population of Interest
  - Person
  - Place
  - Time

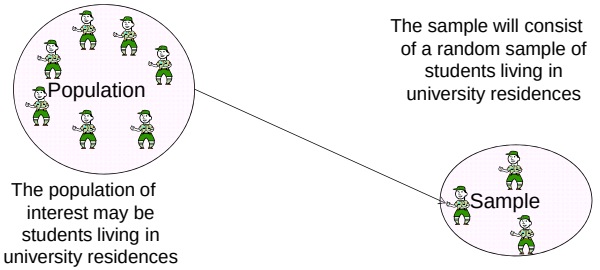
Unit of Analysis

- People
- Cases: Business/Company
- Data: customer transactions

Sample & sampling

- **Sample:**
  A subset of the research population by which participants in your study are selected
- **Sampling**:
  Sampling is the process of selecting a portion of the population, in the research area, which will be a representation of the whole population.

Sampling



Sampling strategy & Frame

- **Sampling strategy:**
  The sampling strategy is the plan that you use to select participants to be sure that the sample represents the population from which you drew your sample
- **Sampling Frame**
  A list of members in the research population
- **Sampling Unit**

More terminology

- **Sampling bias:**
  Sampling bias occurs when the *units* that are selected from the *population* for inclusion in the *sample* are *not characteristic* of (i.e. do not reflect) the *research population*. This can lead to your *sample* being *not representative* of the *research population*.
- **Sampling error:**
  When there is a fluctuation of the statistical value from one sample to another when it is calculated from the same research population.

Why sampling?

- Financial constraints
- Time constraints

Two groups of sampling strategies

- Non-probability sampling techniques
  Non-probability sampling techniques rely on the *subjective judgement* of the researcher when *selecting units* from the *population* to be included in the *sample*
- Probability sampling techniques
  Probability sampling techniques use *random selection* (i.e. *probabilistic methods*) to *select units* from a *sampling frame* (i.e. *similar* or *exactly the same* as the *research population*) to be included in the *sample*.

Rationale for using non-probability sampling strategies

- *Non-probability* sampling requires researchers to use their *subjective judgements*, drawing on *theory* and *the experience of the researcher* and the *evolutionary nature* of the research process.
- Goal is *not objectivity, representativity or generalisability*; rather interest is in intricacies of sample.
- Quicker, easier, cheaper
- Useful in *exploratory* research – to determine if an issue exists

Non-Probability Sampling Strategies

- Convenience or Haphazard Sampling
  easiest to access
- Consecutive Sampling
  seeks to include *all* accessible subjects as part of the sample
- Purposive Sampling – Judgmental, selective or subjective sampling
  rely on the *judgement* of the researcher
- Quota Sampling
  proportional quota sampling
- Volunteer Sampling - Self-selection sampling
  research subjects (or organisations) *volunteer* to take part in the research
- Snowball Sampling
  when the *population* is *hidden* and/or *hard-to-reach*
- Targeted Sampling – Judgment sampling
  A *catch-all.* Participants are selected for a specific feature, treatment, or condition

Rationale for using probability sampling strategies

- To make statistical inferences
  to generalise results from the sample to the research population
- To achieve a representative sample
  to generalise results from the sample to the research population
- To minimise sampling bias

Probability Sampling Strategies

- Simple Random Sampling(not haphazard)
  There is an *equal chance* (*probability*) that each of the units could be selected for inclusion in the sample
- Stratified Random Sampling
  If we are interested in particular *strata* (meaning *groups*) within the *population*
- Systematic Random Sampling
  One typically uses *random number tables* to select the *first unit* for inclusion in the sample, the *remaining units* are selected in an *ordered way* (e.g. every 9th student).
- Cluster Random Sampling (sample size should be at least double)
  Divide population into clusters; select random clusters
- Multi-Stage Area Random Sampling
  A complex form of cluster sampling. The sample is selected in stages, often taking into account the hierarchical (nested) structure of the population.

Sample size

One day there was a fire in a wastebasket in the Dean's office and in rushed a physicist, a chemist, and a statistician. The physicist immediately starts to work on how much energy would have to be removed from the fire to stop the combustion. The chemist works on which reagent would have to be added to the fire to prevent oxidation. While they are doing this, the statistician is setting fires to all the other wastebaskets in the office. "What are you doing?" they demanded. "Well to solve the problem, obviously you need a large sample size" the statistician replies.

Hugh Foley.

Sample size

- Small populations – no need to sample
- Sample size need not be a specific percentage of the study population
- Use statistical formulae – using information from researcher
- Statistical significance may not be *epidemiological* significance i.e. statistical significance may not be significance within the research discipline that the quantitative study is done.

Considerations in calculating sample size

- Statistical power
  The probability that a treatment effect will be detected if it is there. Usually set at 0.8 (80). – pre-determined
- P-level
  The probability of detecting a statistically significant difference that is the result of chance, not the result of the treatment – pre-determined
- Treatment variability
  Influence of treatment on measured responses.
- Error variability
  "Error" contributes to the variability of measured responses other than those that result from the treatment that could alter these measured responses / which includes the inconsistency inherent in measurements obtained with a measurement device or technique that is not perfect

E.G. - Sample size calculation

To *estimate* a single proportion one needs to specify

- Anticipated population proportion
- Level of confidence – 95, 99, etc
- Acceptable margin of error – desired precision

Sample size calculation

To *test* for differences between 2 population parameters, one needs to specify

- Level of significance: α = 0.05, 0.01, etc
- Power of the test: 1- β = 0.8 (80)
- Anticipated outcomes in the group(s) e.g. incidence or means, standard deviations in group(s)
- For estimation also specify effect size (*difference worth finding)*

Other factors

- If interested in more than one outcome – calculate sample size for rarest outcome
- Sample size should be larger to allow for non-response
- Larger sample sizes if not simple random sampling
- Estimate the power of the study under assumptions
- In case-control can have more than one control per case

N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S
10	10	160	113	310	172	460	210	610	236	760	255	910	270	1100	285	2600	335	8500	368	50000	381	170000	383	470000	384	1200000	384	1950000	384
20	19	170	118	320	175	470	212	620	237	770	257	920	271	1200	291	2700	336	9000	368	55000	381	190000	383	500000	384	1250000	384	2050000	384
30	28	180	123	330	178	480	214	630	239	780	258	930	272	1300	297	2800	338	9500	369	60000	382	210000	383	550000	384	1300000	384	2100000	384
40	36	190	127	340	181	490	216	640	240	790	259	940	273	1400	302	2900	339	10000	370	65000	382	230000	384	600000	384	1350000	384	2150000	384
50	44	200	132	350	183	500	217	650	242	800	260	950	274	1500	306	3000	341	10500	371	70000	382	250000	384	650000	384	1400000	384	2200000	384
60	52	210	136	360	186	510	219	660	243	810	261	960	275	1600	310	3500	346	11000	371	75000	382	270000	384	700000	384	1450000	384	2250000	384
70	59	220	140	370	189	520	221	670	244	820	262	970	275	1700	313	4000	351	11500	372	80000	382	290000	384	750000	384	1500000	384	2300000	384
80	66	230	144	380	191	530	223	680	246	830	263	980	276	1800	317	4500	354	12000	372	85000	382	310000	384	800000	384	1550000	384	2350000	384
90	73	240	148	390	194	540	225	690	247	840	264	990	277	1900	320	5000	357	15000	375	90000	383	330000	384	850000	384	1600000	384	2400000	384
100	80	250	152	400	196	550	226	700	248	850	265	1000	278	2000	322	5500	359	20000	377	95000	383	350000	384	900000	384	1650000	384	2450000	384
110	86	260	155	410	199	560	228	710	250	860	266	1010	278	2100	325	6000	361	25000	378	100000	383	370000	384	950000	384	1700000	384	2500000	384
120	92	270	159	420	201	570	230	720	251	870	267	1020	279	2200	327	6500	363	30000	379	105000	383	390000	384	1000000	384	1750000	384	2550000	384
130	97	280	162	430	203	580	231	730	252	880	268	1030	280	2300	329	7000	364	35000	380	110000	383	410000	384	1050000	384	1800000	384	2600000	384
140	103	290	165	440	205	590	233	740	253	890	269	1040	281	2400	331	7500	365	40000	381	130000	383	430000	384	1100000	384	1850000	384	2650000	384
150	108	300	169	450	207	600	234	750	254	900	269	1050	281	2500	333	8000	367	45000	381	150000	383	450000	384	1150000	384	1900000	384	2700000	384

precision =	0.0500
normal prob =	0.975
z =	1.960
p =	0.5
q =	0.5
Z^2pq =	0.9604
Width of interval =	10.00

Infinite Population size: n = 384.1

Explaining sample size

Attempting a more practical example, a professor explained the scientific rigor and meticulous sample selection of the Nielsen television ratings which are used to determine how multiple millions of advertising dollars are spent. A student remained unimpressed saying, “You mean that just a sample of a few thousand can tell us exactly what over 250 MILLION people are doing?”

Finally, the professor, somewhat disgruntled with the scepticism, replied, “Well, the next time you go to the campus clinic and they want to do a blood test…tell them that’s not good enough …tell them to TAKE IT ALL!!”

Reference

Joubert, G. & Ehrlich, R. (eds.) 2007. *Epidemiology: A research manual for South Africa,* Cape Town: Oxford University Press.

Kandace J. Landreneau, “Sampling Strategies”.
http://www.natco1.org/research/files/SamplingStrategies.pdf